MORPHSTATE: COUPLING STATE AGGREGATION AND DETAIL RECOVERY FOR FINE-GRAINED SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

State space models represented by Mamba achieve efficient long-range modeling through selective state updates. However, for visual models that aggregate spatial features into a small number of states, deciding which information should be written into the states does not directly reveal which local structures remain insufficiently represented, leaving subsequent detail recovery without an explicit reference to the current aggregation. To address this issue, we propose MorphState, a fine-grained segmentation framework that uses the spatial allocation of state writes as a shared reference for global aggregation and for generating recovery candidates through regularized feature reconstruction. First, high-resolution morphological cues modulate the spatial write weights while keeping the state count fixed. Next, the updated allocation is treated as a spatial reconstruction dictionary to regularize the reconstruction of pre-aggregation features, and the components that cannot be sufficiently explained by this dictionary are extracted as complementary candidates. Finally, because such residual components may also contain irrelevant texture variations, the model combines morphological orientation, residual magnitude, and coarse semantic predictions to modulate their class-specific contributions and refine high-resolution segmentation outputs. On BD3-Seg, MDMCS, and CUBIT-Seg, MorphState achieves a five-class mIoU of 85.77%, and foreground mIoUs of 83.45% and 53.07%, respectively, outperforming the strongest compared methods on the three datasets by 5.00, 4.32, and 4.18 percentage points. Ablations show additional gains from the complete allocation-and-recovery configuration over MGCR alone on BD3-Seg and MDMCS.

1 INTRODUCTION

Fine-grained segmentation must determine both the region to which a pixel belongs and the local changes that define a region's boundary and connectivity. On building surfaces, thin cracks may occupy only a few pixels, whereas spalling and corrosion require recognition of larger regions. Stone texture, material interfaces, and illumination changes can produce similar local gradients. Visual inspection datasets differ in defect categories and annotation levels (Yang et al., 2022). BD3, MDMCS, and CUBIT-Seg pose damage-segmentation tasks with these distinct morphologies(Xu et al., 2026; Guo et al., 2026; CUHK-USR-Group, n.d.). The challenge is to give sparse, easily disturbed structural evidence representation commensurate with its value for prediction.

These morphological differences require a model to exploit global context to distinguish damage from visually similar textures while preserving the local variations that define object boundaries and connectivity. State space models (SSMs) provide an efficient way to capture such context. Mamba models long-range dependencies through input-dependent selective state updates, with computational cost scaling linearly with sequence length (Gu & Dao, 2024). Vision Mamba extends this principle to visual representation learning through bidirectional state propagation, enabling contextual interaction across image patches (Zhu et al., 2024). For lightweight visual modeling, EfficientViM further aggregates spatial features into a small set of shared states and performs channel mixing in the compressed state space, thereby reducing the cost of global interaction (Lee et al., 2025). However, efficiently forming a global representation does not by itself ensure that the local evidence required for fine-grained segmentation is adequately represented. In models based on shared-state

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

aggregation, state allocation determines which combinations of spatial features are written into the states, but does not explicitly reveal which local variations remain insufficiently represented and therefore require recovery. Fine-grained segmentation thus requires not only adapting the spatial support of a fixed number of states to thin structures and irregular regions, but also identifying local evidence that is complementary to the aggregation actually used by the model.

Prior work improves dense prediction through local-detail modeling, cross-scale fusion, or compensation for state compression(Yu et al., 2021; Xu et al., 2023; Fu et al., 2025; Zhu et al., 2026). These approaches highlight the value of local recovery, but providing local features does not by itself specify their complementarity to the current aggregation. Our key observation is that the spatial patterns used to write features into states can also serve as a dictionary for reconstructing the pre-aggregation features. A reconstruction residual under this dictionary therefore defines recovery candidates relative to the actual allocation. These candidates may contain both meaningful structures and irrelevant textures; reconstruction alone does not determine their value for segmentation.

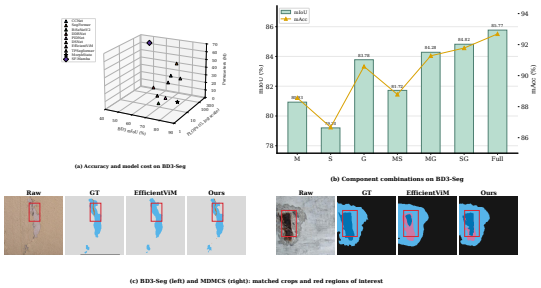


Figure 1: Motivation. (a) BD3-Seg accuracy and model cost (log FLOPs); (b) component ablations; (c) selected BD3-Seg and MDMCS predictions.

Based on this observation, we propose MorphState, which links state aggregation and high-resolution recovery through a shared allocation. Morphology-Aligned State Allocation (MASA) uses high-resolution morphological cues to adapt state support to thin structures and irregular regions while keeping the state count fixed. State-Span Complement Recovery (SSCR) then performs ridge-regularized reconstruction under the updated allocation to extract complementary residuals. A Morphology-Guided Class Refinement (MGCR) decoder combines these residuals with shallow spatial detail and uses the shared morphology field to guide directional feature mixing. It further combines morphological cues, relative residual magnitudes, and coarse semantic posteriors to generate class-specific gates that modulate residual logit corrections to the semantic prediction. The allocation thus defines the reference for recovery, while the decoder determines how the resulting evidence refines each class prediction.

Our contributions are threefold. First, we use state allocation as a shared spatial reference for aggregation and recovery: the allocation used to write features into compact states also serves as a regularized reconstruction dictionary for generating recovery candidates. MASA adapts this reference using morphological evidence without increasing the state count. Second, MGCR combines directional feature mixing with class-specific gates to turn recovery candidates into prediction corrections. Third, MorphState improves mean IoU over the strongest evaluated baselines on three infrastructure-damage datasets. Ablations examine alternative allocation and recovery designs and show additional gains from the complete configuration over MGCR alone on BD3-Seg and MDMCS.

2 Related Work

Visual state space models State space models (SSMs) enable long-range modeling through structured state representations. S4 uses structured parameterization for efficient sequence modeling, while Mamba introduces input-dependent selective updates (Gu et al., 2022; Gu & Dao, 2024). State Space Duality (SSD) connects SSMs and attention through structured matrix representations and underlies Mamba-2 (Dao & Gu, 2024). For vision, Vision Mamba uses bidirectional state-space modeling, VMamba aggregates context along four scanning routes, and LocalMamba strengthens local dependencies through windowed scans (Zhu et al., 2024; Liu et al., 2024; Huang et al., 2025). EfficientVMamba combines atrous selective scanning with a convolutional branch for lightweight global and local modeling (Pei et al., 2025). Spatial-Mamba introduces structure-aware fusion of neighboring hidden states (Xiao et al., 2025). VSSD develops a noncausal SSD formulation, while EfficientViM uses a hidden state mixer (HSM) to perform channel mixing in compressed hidden states (Shi et al., 2025; Lee et al., 2025). SF-Mamba uses patch swapping for bidirectional informa-

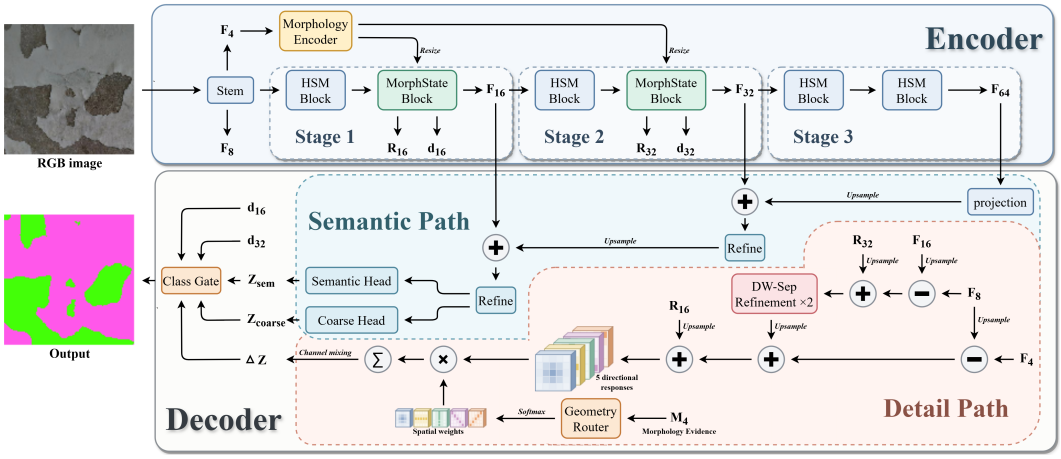


Figure 2: Overall MorphState architecture. The encoder uses morphological evidence in the last block of each of its first two state stages to adjust state writes and extract residuals complementary to the current allocation. The decoder fuses semantics with high-resolution detail and corrects predictions through directional morphology mixing and class-specific gating.

tion flow and batch folding with periodic state resets for GPU parallelism (Yoshimura et al., 2026). Building on EfficientViM's HSM-SSD formulation, MorphState uses high-resolution morphological evidence to adapt spatial state allocation at a fixed state count.

Structure-aware dense prediction and complementary recovery Dense prediction models combine context and local detail through encoder–decoder fusion, high-resolution streams, hierarchical features, attention, and multi-branch exchange (Chen et al., 2018; Wang et al., 2021; Xie et al., 2021; Huang et al., 2019; Yu et al., 2021; Pan et al., 2023; Xu et al., 2023). Boundary-focused methods use shape streams, adaptive point refinement, or interior-to-boundary label transfer (Takikawa et al., 2019; Kirillov et al., 2020; Yuan et al., 2020b); TPSegformer preserves spatial continuity in building-defect decoding (Xu et al., 2026). OCR's soft object regions motivate our within-image semantic anchors (Yuan et al., 2020a). Among SSM-based methods, SegMAN combines local attention with multiscale context (Fu et al., 2025), while RS-SSM uses spectral cues for complementary spatiotemporal recovery (Zhu et al., 2026). MorphState instead defines complementary evidence relative to the current state-allocation subspace and uses morphology-guided, class-specific refinement for high-resolution prediction.

3 Method

MorphState organizes dense prediction with a limited number of states into three connected steps: adjusting state writes using morphological evidence, constructing complementary candidates from the actual allocation subspace, and determining their contribution to prediction using class semantics. The first two steps share the allocation matrix: MASA modulates it using morphological evidence, and SSCR uses its spatial patterns as a reconstruction dictionary. Figure 2 shows the overall architecture.

3.1 Architecture and State Writes

Given an image $I \in \mathbb{R}^{3\times H\times W}$, we retain the convolutional stem and three-stage hierarchy of EfficientViM-M3. The stem provides shallow features F_4, F_8 at strides 4 and 8, and the state stages output F_{16}, F_{32}, F_{64}. Their channel widths are $(224, 320, 512)$, each stage has two blocks, and their state counts are $(49, 25, 9)$. Morphology-Aligned State Allocation (MASA) and State-Span Complement Recovery (SSCR) operate in the last block of each of the first two stages; Morphology-Guided Class Refinement (MGCR) corrects predictions at stride 4.

Omitting the batch dimension, let $X \in \mathbb{R}^{C\times L}$ denote features after the first local mixing operation in a block and before state mixing, let $\widetilde{X} = \mathrm{LN}(X)$, and let N be the state count. The original HSM writes and reads states as

$$S^0 = \mathrm{softmax}_L(\Delta + a\mathbf{1}_L^\top) \odot B, \qquad H_s = \widetilde{X}(S^0)^\top, \qquad Y = \Phi(H_s)C_r. \tag{1}$$

Here $B, \Delta, C_r \in \mathbb{R}^{N\times L}$ are input-dependent. The backbone parameter $a \in \mathbb{R}^N$ contains one scalar per state, and $\mathbf{1}_L \in \mathbb{R}^L$ is the all-ones vector; $a\mathbf{1}_L^\top$ broadcasts each scalar across spatial positions. Φ denotes gating and channel projection within hidden states. Because B can be negative, S^0 is a *signed state–space allocation*, not a probability matrix. Figure 3 contrasts the original HSM-SSD layer with a MorphState layer: MASA adjusts S^0 in Eq. (1), changing how spatial evidence is written into hidden states, while SSCR reads X and the same adjusted allocation S to construct recovery candidates.

3.2 Morphology-Aligned State Allocation

Shared morphological evidence. We project F_4 to a scalar field u with a learned 1×1 convolution and apply fixed Sobel operators to obtain g_x, g_y. For Gaussian scales $s \in \{1, 2, 3\}$, we construct structure tensors

$$J_s = G_s * \begin{bmatrix} g_x^2 & g_x g_y \\ g_x g_y & g_y^2 \end{bmatrix}, \quad t_s = J_s^{xx} + J_s^{yy}, \quad d_s = \sqrt{(J_s^{xx} - J_s^{yy})^2 + 4(J_s^{xy})^2 + \epsilon}. \tag{2}$$

Each tensor yields a four-channel field describing energy, coherence, and unsigned orientation:

$$M_s = \left[\log(1 + t_s),\ \frac{d_s}{t_s + \epsilon},\ \frac{J_s^{xx} - J_s^{yy}}{d_s},\ \frac{2J_s^{xy}}{d_s}\right], \qquad M = \sum_s \pi_s M_s, \tag{3}$$

Here $\pi = \mathrm{softmax}_s(f_{\mathrm{scale}}([M_1, M_2, M_3]))$ gives pixelwise scale weights. Let E denote fused energy, let κ denote fused coherence clipped to $[0, 1]$, and set $\bar{E} = \max(E, 0)/(\mathrm{mean}_{xy} \max(E, 0) + \epsilon)$. We define three explicit support bases:

$$b_{\mathrm{line}} = 1 + \kappa\bar{E}, \qquad b_{\mathrm{region}} = 1 + (1 - \kappa)\bar{E}, \qquad b_{\mathrm{neutral}} = 1. \tag{4}$$

They provide references for anisotropic structure, more isotropic structure, and a neutral support, respectively; none is assigned a semantic class in advance. We construct the bases at stride 4 and bilinearly interpolate them to the relevant state stages. The raw coherence in Eq. (3) need not lie strictly in the unit interval because of stabilizing terms. Support construction and the class-specific correction gate use the clipped value, while directional recovery uses the original fused field M.

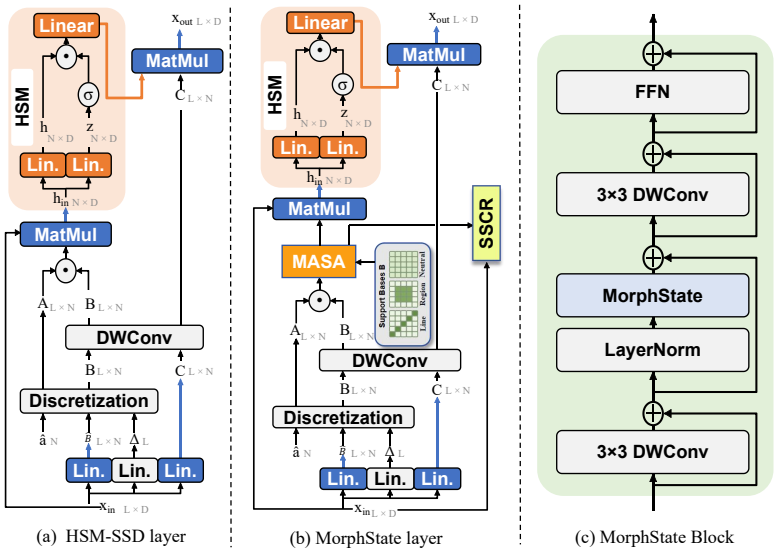
Preferences from existing state coverage. Pretrained states have no fixed linear or areal identity. For state j, MASA measures overlap between its original absolute allocation and each support, then combines the overlap with a learned prior β_{jk} to obtain sample-dependent support preferences:

$$\begin{aligned} e_{jk} &= \frac{\sum_l |\,\mathrm{sg}(S_{jl}^0)|\,\mathrm{sg}(b_{kl})}{\sum_l |\,\mathrm{sg}(S_{jl}^0)|}, \\ p_{jk} &= \mathrm{softmax}_k\left(\beta_{jk} + \eta \log \frac{e_{jk}}{\frac{1}{3}\sum_{k'} e_{jk'}}\right), \qquad q_{jl} = \sum_k p_{jk}\,\mathrm{sg}(b_{kl}). \end{aligned} \tag{5}$$

Here sg denotes stop-gradient, β is initialized to zero, and $\eta = 1$; we omit ϵ clamps in denominators and logarithms for brevity. The evidence branch observes the current allocation and morphological supports, while S^0 on the final allocation path remains differentiable. The normalized entropy of support preferences defines confidence $c_j = 1 + \sum_k p_{jk} \log p_{jk}/\log 3$.

Conservative reallocation of absolute allocation mass. Define $\langle v\rangle_j = \sum_l |S_{jl}^0| v_{jl} / \sum_l |S_{jl}^0|$. We first remove the overall support scale, then bound logarithmic reallocation and renormalize:

$$\begin{aligned} u_{jl} &= \frac{q_{jl}}{\langle q\rangle_j}, \qquad r_{jl} = \exp\left[\delta \tanh\left(\frac{\log u_{jl}}{\delta}\right)\right], \\ S_{jl}^g &= S_{jl}^0 \frac{r_{jl}}{\langle r\rangle_j}, \qquad S_{jl} = S_{jl}^0 + \sigma(\gamma_j) c_j (S_{jl}^g - S_{jl}^0). \end{aligned} \tag{6}$$



(a) HSM-SSD layer (b) MorphState layer (c) MorphState Block

Figure 3: State-block computation. Left: the original HSM-SSD layer. Middle: a MorphState layer in which MASA adjusts state allocation and SSCR extracts complementary candidates from the current allocation. Right: a residual block comprising local depthwise convolutions, a MorphState layer, and a feed-forward network.

By default, $\delta = 0.5$ and $\gamma_j = -4$, keeping the initial update close to the original allocation. Positive modulation and within-state normalization ensure $\operatorname{sign}(S_{jl}) = \operatorname{sign}(S^0_{jl})$ and $\sum_l |S_{jl}| = \sum_l |S^0_{jl}|$, up to floating-point error and clamps on very small denominators. This constraint preserves each state's *absolute allocation mass*, not its response or information content. The number of states is unchanged.

3.3 State-Span Complement Recovery

State reallocation determines which spatial combinations enter compact states, but does not ensure that they express all pixel-level evidence. SSCR takes the *row space of the current allocation* as a coverage reference and reconstructs the pre-mixing features X with regularization. We first normalize each row as $\bar{S}_{j,:} = S_{j,:}/\sqrt{\|S_{j,:}\|_2^2 + \epsilon}$ and then solve

$$A^* = \arg\min_A \|X - A\bar{S}\|_F^2 + \lambda\|A\|_F^2, \qquad R = X - (X\bar{S}^\top)(\bar{S}\bar{S}^\top + \lambda I_N)^{-1}\bar{S}. \tag{7}$$

R contains feature candidates insufficiently explained by the current state subspace. We reconstruct X itself, *not* $X - Y$: the hidden-state response Y has also undergone nonlinear channel mixing and a separate readout, so a response difference does not isolate deficient allocation coverage. The default implementation does not stop gradients through S in SSCR.

We use $\lambda = 10^{-4}$ by default. Because $\lambda > 0$, Eq. (7) yields a regularized residual rather than an exact orthogonal complement. If s_i is a nonzero singular value of $\bar{S}$, the reconstruction operator has weight $s_i^2/(s_i^2 + \lambda)$ in the corresponding spatial direction. It continuously suppresses components explained by the allocation subspace while retaining other candidates. This interpretation does not require Y to equal a linear projection of X; subsequent morphological and semantic conditions determine the task relevance of R.

We further define a relative residual magnitude at each location:

$$d_l = \frac{\sqrt{\frac{1}{C}\sum_c R_{cl}^2 + \epsilon}}{\sqrt{\frac{1}{C}\sum_c X_{cl}^2 + \epsilon} + \epsilon}. \tag{8}$$

R provides recovery content, while d supplies the later gate with a cue to relative undercoverage. The quantity d is neither an error probability nor constrained to $[0, 1]$. We solve an $N \times N$ linear

system in FP32 without forming an $L \times L$ projection matrix. Complexity is $O(CNL + N^2L + N^3)$, linear in spatial length L for fixed N.

3.4 Morphology-Guided Class Refinement

Roles of semantics and complementary evidence. The semantic branch projects F_{16}, F_{32}, F_{64} to 192 channels and uses local multi-dilation context and a top-down pyramid to produce stride-16 semantic features T_{16}, which are interpolated to T_4. The recovery branch has width 96. Let P_s, Q_s denote independent channel projections and U_s interpolation to stride s. We form shallow cross-scale differences and inject SSCR residuals from two levels:

$$V_8 = \Psi_8(P_8 \cdot F_8 - U_8(P_{16} \cdot F_{16}) + U_8(Q_{32} \cdot R_{32})), \tag{9}$$

$$V_4 = P_4 \cdot F_4 - U_4(P_8 \cdot F_8) + U_4(V_8) + U_4(Q_{16} \cdot R_{16}). \tag{10}$$

These differences encode contrasts between independently projected features at adjacent scales and combine with the state-subspace residuals to form recovery features. The shared morphology field predicts pixelwise weights $w = \mathrm{softmax}(f_{\mathrm{dir}}(M))$ for five depthwise-convolution branches: isotropic, horizontal, vertical, and two diagonal orientations:

$$D_4 = \Psi_{\mathrm{mix}}\left(\sum_{a=1}^{5} w_a(M) \odot K_a(V_4)\right). \tag{11}$$

The isotropic branch uses a 3×3 kernel; the oriented branches have length 7. Fixed diagonal masks on convolution kernels implement the diagonal branches. Direction weights are learned, without assigning a required direction to any class.

Within-image semantic anchors. The coarse head produces $Z_c = h_c(T_{16})$; its interpolated, stop-gradient posterior is $P_c = \mathrm{sg}(\mathrm{softmax}(U_4 Z_c))$. We use P_c to softly pool channel-normalized T_4, then normalize the result to obtain image-specific class prototypes a_k. The semantic prediction is

$$\begin{aligned} a_k &= \mathrm{norm}\left(\frac{\sum_l P_{c,kl}\, \mathrm{norm}(T_{4,l})}{\max(\sum_l P_{c,kl}, \epsilon)}\right), \\ Z_{\mathrm{sem},kl} &= h_{\mathrm{sem}}(T_4)_{kl} + \frac{\alpha}{\tau}\, \mathrm{norm}(T_{4,l})^\top a_k. \end{aligned} \tag{12}$$

The temperature is $\tau = 0.15$, and the learned scalar α is initialized to zero. This auxiliary calibration maintains an independent prediction capability in the semantic branch before recovery.

Class-specific residual correction. Recovery content need not affect every class equally. Let K be the number of semantic classes. MGCR predicts class-specific residual logits ΔZ from D_4 and combines the full coarse posterior, morphology, and two coverage cues into a K-channel gate:

$$\begin{aligned} \Omega &= \sigma\big(f_{\mathrm{gate}}([P_c, \bar{E}, \kappa, U_4 d_{16}, U_4 d_{32}, \mathcal{H}(P_c)])\big), \\ \Delta Z &= h_{\mathrm{detail}}(D_4), \qquad Z = Z_{\mathrm{sem}} + \Omega \odot \Delta Z. \end{aligned} \tag{13}$$

Here $\mathcal{H}(P_c) = -\sum_k P_{c,k} \log P_{c,k} / \log K$. We initialize both weights and biases of h_{detail} to zero, so the initial prediction exactly equals the semantic branch; the gate's last-layer weights are zero-initialized with bias -1. Finally, Z is bilinearly interpolated to input resolution. Class-specific gates model distinct contributions of the same detail instead of controlling the injection of all semantic features with one pixelwise scalar.

3.5 Training Objective

The final prediction and coarse semantic head use cross-entropy over valid labeled pixels:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}}(U_1 Z, Y_{\mathrm{gt}}) + 0.1\mathcal{L}_{\mathrm{CE}}(U_1 Z_c, Y_{\mathrm{gt}}). \tag{14}$$

Auxiliary supervision of the coarse head provides explicit semantic conditions for posterior gating and semantic anchors. Training requires neither additional boundary labels nor subspace-reconstruction supervision. The learned morphology projection, state allocation, and recovery branch are optimized jointly for segmentation.

4 EXPERIMENTS

4.1 DATASETS AND IMPLEMENTATION DETAILS

We train and evaluate separate models on BD3-Seg, MDMCS (Guo et al., 2026), and CUBIT-Seg (Yang et al., 2025; CUHK-USR-Group, n.d.), using train/validation/test splits of 500/100/100, 1000/100/100, and 4635/993/994 images, respectively. The primary metric is five-class mIoU on BD3-Seg and foreground mIoU on MDMCS and CUBIT-Seg. All metrics are computed from test-set confusion matrices. Foreground mIoU excludes Background from class averaging, but still counts foreground false positives on background pixels. For BD3-Seg, original zero labels and padding pixels are ignored. For MorphState, we initialize the backbone with the official distilled EfficientViM-M3 weights (Lee et al., 2025) and use 512×512 inputs. BD3-Seg and MDMCS are resized with aspect-ratio-preserving padding, whereas CUBIT-Seg is resized directly. For MorphState, we use AdamW (Loshchilov & Hutter, 2019) with batch size 8, weight decay 0.03, and learning rates of 6×10^{-4} and 6×10^{-5} for new and pretrained parameters, respectively. A 10-epoch warm-up is followed by polynomial learning-rate decay. Training uses mixed precision and an exponential moving average (EMA) with decay 0.999. Checkpoints are selected by validation mIoU over all classes and then evaluated on the test set. All comparison methods use the same fixed train/validation/test splits, class-label definitions, and original-resolution test-set metrics for each dataset.

4.2 COMPARISON WITH REPRESENTATIVE METHODS

In Tables 1–3, bold and underline mark the best and second-best value in each column. The class columns report IoU; method references and venues appear in Table 1. BD3-Seg averages over five valid classes; MDMCS foreground mIoU excludes Background, while its mAcc includes it; CUBIT-Seg foreground averages exclude Background.

Table 1: BD3-Seg test results (%).

Method	Venue/Year	Fiss.	Peel	Moss	Lime	Stone	mIoU	mAcc
CCNet (2019)	ICCV 2019	53.86	81.21	60.03	94.83	70.90	72.17	82.53
SegFormer (2021)	NeurIPS 2021	50.60	72.49	57.35	95.99	**93.72**	74.03	86.43
BiSeNetV2 (2021)	IJCV 2021	45.11	65.43	58.70	94.43	78.00	68.33	77.50
DDRNet (2023)	T-ITS 2023	48.68	44.99	54.95	87.81	78.36	62.96	88.49
PIDNet (2023)	CVPR 2023	55.90	78.60	55.37	95.07	79.71	72.93	84.27
DSNet (2025)	TCSVT 2025	48.17	64.73	32.81	93.22	78.73	63.53	80.91
EfficientViM (2025)	CVPR 2025	<u>62.87</u>	89.26	63.93	95.17	77.38	77.72	85.36
SF-Mamba (2026)	ICML 2026	38.52	34.40	37.97	80.74	46.32	47.59	66.84
TPSegformer (2026)	CVPR 2026	59.98	**91.11**	<u>72.58</u>	**96.71**	<u>83.44</u>	<u>80.77</u>	<u>90.22</u>
MorphState	–	**75.40**	<u>90.36</u>	**83.65**	<u>96.27</u>	83.16	**85.77**	**92.69**

Table 2: MDMCS test results (%).

Method	Bkg.	Crack	Spall.	Corr.	Rebar	$\mathrm{mIoU}_{\mathrm{fg}}$	mAcc
CCNet	97.72	57.95	89.57	<u>80.89</u>	67.56	73.99	<u>92.03</u>
SegFormer	97.79	75.20	90.53	77.43	<u>73.36</u>	<u>79.13</u>	**93.52**
BiSeNetV2	92.30	47.55	70.02	69.74	56.80	61.02	88.92
DDRNet	95.52	68.83	78.94	75.04	55.59	69.60	86.22
PIDNet	94.95	63.29	78.70	71.94	58.38	68.08	89.38
DSNet	92.23	50.58	68.80	69.46	52.29	60.28	87.22
EfficientViM	<u>98.25</u>	63.00	<u>91.23</u>	80.15	63.31	74.42	83.85
SF-Mamba	97.19	<u>79.27</u>	85.99	72.10	51.13	72.12	85.96
TPSegformer	97.93	62.98	90.87	77.82	67.30	74.75	89.70
MorphState	**98.61**	**81.11**	**93.02**	**85.08**	**74.57**	**83.45**	91.45

Table 3: CUBIT-Seg test results (%).

Method	Bkg.	Crack	Spall.	$\mathrm{mIoU}_{\mathrm{fg}}$	$\mathrm{mAcc}_{\mathrm{fg}}$	mDice
CCNet	<u>95.31</u>	<u>43.83</u>	<u>53.95</u>	<u>48.89</u>	**65.42**	<u>65.52</u>
SegFormer	94.30	34.99	44.01	39.50	52.17	56.48
BiSeNetV2	92.96	33.45	38.04	35.75	53.05	52.63
DDRNet	94.18	30.98	41.21	36.09	48.06	52.84
PIDNet	93.45	31.98	33.84	32.91	45.15	49.52
DSNet	92.32	26.90	34.07	30.49	47.18	46.61
EfficientViM	95.18	40.73	39.09	39.91	52.43	57.05
SF-Mamba	94.42	39.11	37.95	38.53	46.98	55.62
TPSegformer	95.20	41.99	41.10	41.55	56.19	58.70
MorphState	**96.00**	**47.38**	**58.77**	**53.07**	<u>65.18</u>	**69.16**

BD3-Seg. Table 1 shows that MorphState reaches 85.77% mIoU, exceeding TPSegformer by 5.00 points and EfficientViM by 8.05 points. Its Fissures and Moss IoUs are 75.40% and 83.65%, gains of 12.53 and 11.07 points over the strongest comparator for each class. These classes reflect demands for thin-structure and irregular-region segmentation, consistent with the motivation for morphology-aligned allocation and complementary recovery. Comparators retain the best values for Peel off, Lime, and Stone; the aggregate gain does not require every class to improve.

MDMCS. On MDMCS, MorphState achieves 83.45% $\mathrm{mIoU}_{\mathrm{fg}}$, 4.32 points above SegFormer, and the highest IoU in each of the four foreground classes (Table 2). Crack IoU is 81.11%, 1.84 points above SF-Mamba's 79.27%; all-class mAcc is 91.45%, below SegFormer's 93.52%. **CUBIT-Seg.** On CUBIT-Seg, MorphState obtains 53.07% $\mathrm{mIoU}_{\mathrm{fg}}$, 4.18 points above CCNet. Crack and Spalling improve by 3.55 and 4.82 points, respectively (Table 3). Foreground mDice increases by 3.64 points while foreground mAcc is 0.24 points lower, suggesting that the main gain lies in overlap between predicted and ground-truth regions.

Computational cost. Appendix A compares MorphState and TPSegformer under 512×512 inputs, batch size one, and five output classes (Table A.1). MorphState has 15.693 M parameters and an estimated 5.039 G FLOPs, while TPSegformer has 34.631 M parameters and an estimated 28.813 G FLOPs. Relative to the EfficientViM comparison model in the same table, MorphState reduces parameters from 18.695 M to 15.693 M and increases estimated FLOPs from 4.492 G to 5.039 G. On BD3-Seg, SF-Mamba-Tiny has 61.668 M inference parameters and approximately 288.610 G FLOPs. Its custom-operator analytic estimate uses a different counting method, described in Appendix A.

4.3 Ablation Studies

Core component ablations. All ablations on BD3-Seg use the same training and evaluation protocol. As shown in Table 4, MASA+SSCR, MASA+MGCR, and SSCR+MGCR achieve 81.72%, 84.28%, and 84.82% mIoU, respectively, each exceeding either of its constituent single-component configurations. Adding the third component yields 85.77% mIoU for the full model, gains of 4.05, 1.49, and 0.95 percentage points over these three two-component configurations, respectively. Together, the single-component, two-component, and full-model results support the design of linking morphology-aligned aggregation and complementary recovery through shared state allocation, and demonstrate the effectiveness of MGCR in improving high-resolution predictions through directional recovery and class-specific residual correction. Further mechanism ablations are reported in Appendix C.

Table 4: Component ablations on BD3-Seg (%). ✓ indicates an enabled component. Class columns report IoU; bold and underline indicate the best and second-best values in each column, respectively.

MASA	SSCR	MGCR	Fissures	Peel off	Moss	Lime	Stone	mIoU	mAcc
×	×	×	62.87	89.26	63.93	95.17	77.38	77.72	85.36
✓	–	–	64.78	88.35	76.59	95.27	79.64	80.93	88.57
–	✓	–	64.06	89.14	65.02	95.99	81.80	79.20	86.67
–	–	✓	<u>74.36</u>	90.00	75.25	**96.36**	82.94	83.78	90.59
✓	✓	–	66.03	87.61	76.19	95.92	82.87	81.72	88.79
✓	–	✓	73.68	89.36	78.65	96.26	**83.46**	84.28	91.28
–	✓	✓	74.35	<u>90.05</u>	<u>79.91</u>	<u>96.34</u>	<u>83.45</u>	<u>84.82</u>	<u>91.78</u>
✓	✓	✓	**75.40**	**90.36**	**83.65**	96.27	83.16	**85.77**	**92.69**

The component effects differ on MDMCS (Table C.1). MGCR alone achieves 81.29% $\mathrm{mIoU}_{\mathrm{fg}}$; adding MASA or SSCR individually yields 80.73% or 80.93%, while their joint inclusion reaches 83.45%. Thus, individual component gains depend on the dataset and the accompanying modules. The complete configuration achieves the highest aggregate score in both component studies. Appendix C further examines allocation strategies, residual formulations, and the operators within MGCR.

4.4 Qualitative Comparison

Figure 4 shows three BD3-Seg cases illustrating thin-structure preservation, region-boundary recovery, and local class separation. Additional MDMCS and CUBIT-Seg comparisons are in Appendix B (Figures B.1 and B.2). All figures include every model method in the corresponding quantitative result table, alongside the raw image and ground truth (GT). Each case shows the full image above a crop from the same red solid box. The class palette is consistent within each dataset, and predicted labels are unaltered. Cases and crop locations were deliberately selected to illustrate favorable examples and do not represent random samples. Dataset-level quantitative results remain the basis for overall performance claims.

BD3-Seg: thin structures and class boundaries. From top to bottom, Figure 4 shows three cases. The first focuses on a fissure–peel-off interface, where MorphState's thin-fissure prediction is closer to GT. The second shows a narrow extension at the edge of moss. The third shows a fissure within stone: MorphState follows the fissure more closely while retaining the surrounding stone, whereas several baselines fragment the fissure or introduce spurious regions. Together, the cases include all five valid BD3-Seg classes.

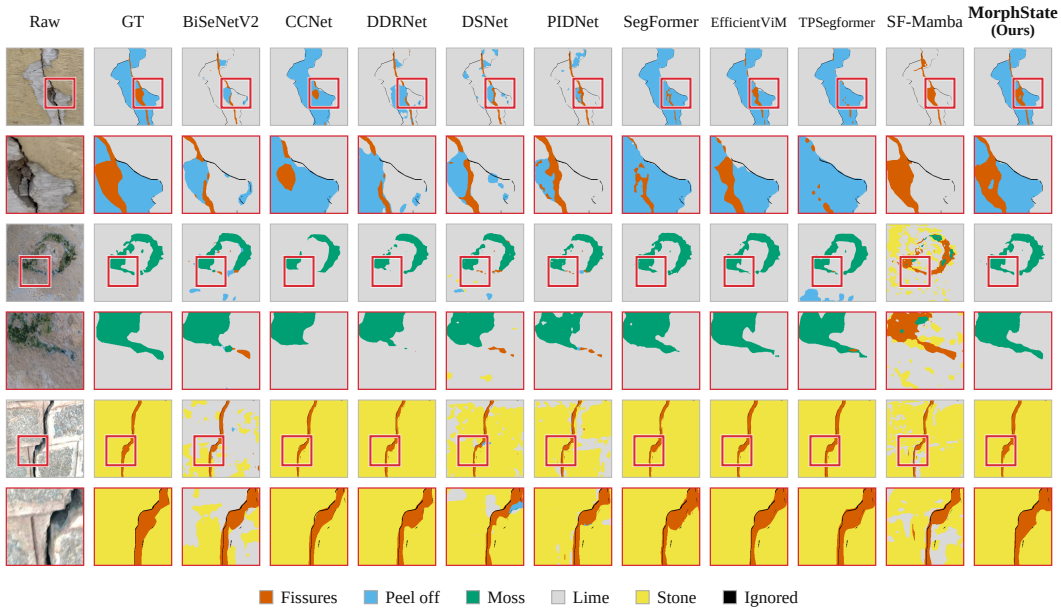


Figure 4: BD3-Seg qualitative comparison of all methods in the quantitative result table. From top to bottom: a fissure–peel-off interface, a moss boundary, and a fissure within stone. Each case shows a full image above the same-location red-box crop. Black denotes ignored pixels, not a background class.

5 Conclusion

We presented MorphState, a state-space framework for fine-grained segmentation that couples compact aggregation with local evidence recovery. Morphological cues adjust spatial state support without increasing the state count, and the updated allocation serves as a reconstruction dictionary for extracting regularized feature residuals. Morphology-guided mixing and class-specific correction then use this complementary evidence to refine high-resolution predictions. Experiments on BD3-Seg, MDMCS, and CUBIT-Seg show improved mean IoU over the evaluated baselines. Component studies identify high-resolution refinement as an important contributor and show additional gains from the complete allocation and recovery configuration, with individual component effects varying across datasets. These findings demonstrate a concrete use of state allocation as a reconstruction reference for generating recovery candidates in infrastructure-damage segmentation.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and polishing, literature retrieval and discovery, and feedback on manuscript clarity and experimental presentation. AI-assisted outputs are treated as suggestions requiring author verification, including checks of factual statements and references. The authors take full responsibility for the final content of this work, including its methods, results, and conclusions.

REFERENCES

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Computer Vision – ECCV 2018*, pp. 833–851. Springer International Publishing, 2018. ISBN 9783030012342. doi: 10.1007/978-3-030-01234-2_49. URL http://dx.doi.org/10.1007/978-3-030-01234-2_49.

CUHK-USR-Group. Defect-Dataset: CUBIT-Seg Dataset and Documentation. Official dataset repository, n.d. URL https://github.com/CUHK-USR-Group/Defect-Dataset. Accessed 2026-09-22.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10041–10071. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/dao24a.html.

Yunxiang Fu, Meng Lou, and Yizhou Yu. SegMAN: Omni-scale Context Modeling with State Space Models and Local Attention for Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19077–19087, June 2025.

Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. In *Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently Modeling Long Sequences with Structured State Spaces. In *The International Conference on Learning Representations (ICLR)*, 2022.

Pengwei Guo, Zhan Jiang, Weina Meng, and Yi Bao. MDMCS: A Benchmark Data Set for Multidamage Monitoring of Concrete Structures. *Journal of Bridge Engineering*, 31(2), February 2026. ISSN 1943-5592. doi: 10.1061/jbenf2.beeng-7893. URL http://dx.doi.org/10.1061/JBENF2.BEENG-7893.

Zilu Guo, Liuyang Bian, Hu Wei, Jingyu Li, Huasheng Ni, and Xuan Huang. DSNet: A Novel Way to Use Atrous Convolutions in Semantic Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(4):3679–3692, 2025. doi: 10.1109/TCSVT.2024.3509504.

Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. LocalMamba: Visual State Space Model with Windowed Selective Scan. In *Computer Vision – ECCV 2024 Workshops*, pp. 12–22. Springer Nature Switzerland, 2025. ISBN 9783031919794. doi: 10.1007/978-3-031-91979-4_2. URL http://dx.doi.org/10.1007/978-3-031-91979-4_2.

Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. CCNet: Criss-Cross Attention for Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. PointRend: Image Segmentation As Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Sanghyeok Lee, Joonmyung Choi, and Hyunwoo J. Kim. EfficientViM: Efficient Vision Mamba with Hidden State Mixer based State Space Duality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14923–14933, June 2025.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual State Space Model. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 103031–103063. Curran Associates, Inc., 2024. doi: 10.52202/079017-3273. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/baa2da9ae4bfed26520bb61d259a3653-Paper-Conference.pdf.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Huihui Pan, Yuanduo Hong, Weichao Sun, and Yisong Jia. Deep Dual-Resolution Networks for Real-Time and Accurate Semantic Segmentation of Traffic Scenes. *IEEE Transactions on Intelligent Transportation Systems*, 24(3):3448–3460, March 2023. ISSN 1558-0016. doi: 10.1109/tits.2022.3228042. URL http://dx.doi.org/10.1109/TITS.2022.3228042.

Xiaohuan Pei, Tao Huang, and Chang Xu. EfficientVMamba: Atrous Selective Scan for Light Weight Visual Mamba. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6):6443–6451, 2025. doi: 10.1609/aaai.v39i6.32690.

Yuheng Shi, Mingjia Li, Minjing Dong, and Chang Xu. VSSD: Vision Mamba with Non-Causal State Space Duality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10819–10829, October 2025.

Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-SCNN: Gated Shape CNNs for Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43 (10):3349–3364, October 2021. ISSN 1939-3539. doi: 10.1109/tpami.2020.2983686. URL http://dx.doi.org/10.1109/TPAMI.2020.2983686.

Chaodong Xiao, Minghan Li, Zhengqiang Zhang, Deyu Meng, and Lei Zhang. Spatial-Mamba: Effective Visual State Space Models via Structure-Aware State Fusion. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b7216f4a324864e1f592c18de4d83d10-Abstract-Conference.html.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf.

Jiacong Xu, Zixiang Xiong, and Shankar P. Bhattacharyya. PIDNet: A Real-Time Semantic Segmentation Network Inspired by PID Controllers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19529–19539, June 2023.

Linkang Xu, Gang Li, Yue Song, and Xiangxin Ji. Hilbert Curve-Based Attention Enabling Topology-Preserving Image Tensor Representation for Semantic Segmentation Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13113–13122, June 2026.

Guidong Yang, Kangcheng Liu, Jihan Zhang, Benyun Zhao, Zuoquan Zhao, Xi Chen, and Ben M. Chen. Datasets and processing methods for boosting visual inspection of civil infrastructure: A comprehensive review and algorithm comparison for crack classification, segmentation, and detection. *Construction and Building Materials*, 356:129226, 2022. doi: 10.1016/j.conbuildmat.2022.129226.

Guidong Yang, Benyun Zhao, Jihan Zhang, Junjie Wen, Qingxiang Li, Lei Lei, Xi Chen, and Ben M. Chen. Det-Recon-Reg: An Intelligent Framework Toward Automated UAV-Based Large-Scale Infrastructure Inspection. *IEEE Transactions on Instrumentation and Measurement*, 74:1–16, 2025. doi: 10.1109/TIM.2025.3571118.

Masakazu Yoshimura, Teruaki Hayashi, Yuki Hoshino, Wei-Yao Wang, and Takeshi Ohashi. SF-Mamba: Rethinking State Space Model for Vision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=R9AUrEgZEq.

Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. BiSeNet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation. *International Journal of Computer Vision*, 129(11):3051–3068, September 2021. ISSN 1573-1405. doi: 10.1007/s11263-021-01515-2. URL http://dx.doi.org/10.1007/s11263-021-01515-2.

Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-Contextual Representations for Semantic Segmentation. In *Computer Vision – ECCV 2020*, pp. 173–190. Springer International Publishing, 2020a. ISBN 9783030585396. doi: 10.1007/978-3-030-58539-6_11. URL http://dx.doi.org/10.1007/978-3-030-58539-6_11.

Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang. SegFix: Model-Agnostic Boundary Refinement for Segmentation. In *Computer Vision – ECCV 2020*, pp. 489–506. Springer International Publishing, 2020b. ISBN 9783030586102. doi: 10.1007/978-3-030-58610-2_29. URL http://dx.doi.org/10.1007/978-3-030-58610-2_29.

Kai Zhu, Zhenyu Cui, Zehua Zang, and Jiahuan Zhou. RS-SSM: Refining Forgotten Specifics in State Space Model for Video Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10741–10752, June 2026.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62429–62442. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhu24f.html.